Introduction to Probability and Statistics

Sec 1: Descriptive Statistics

AIMS | African Institute for Mathematical Sciences CAMEROON

Population and Samples

- **Population:** The set of all units (objects, persons, ...) that we want to draw inference on (*i.e.*, that we ask a question about).
- **Sample:** Subset of the population we collect data on.
- Example: Favourite football team of African university students.
  - Population: Entire set of African students.
  - Sample: The students at AIMS Cameroon this year.
- We analyse samples to make inferences about a population.
- This is **statistical inference** (later in the course).

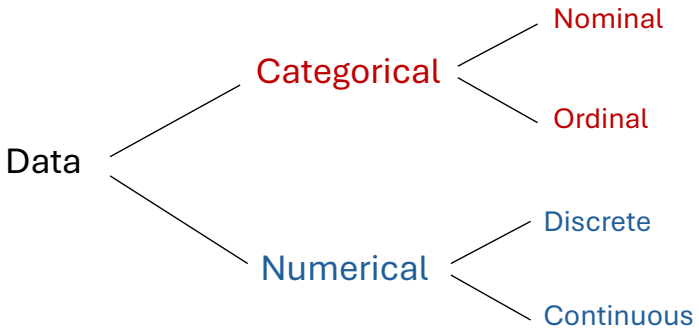
Types of Data

- **Quantitative:** Data values are numerical.
  - Discrete: Only a finite number of values possible
    Shoe size; Number of siblings (brothers/sisters); Number of beds in a house
  - Continuous: Values fall into a continuous range
    Length of a movie; Weight of Person; Air Temperature

- **Categorical:** Data are descriptive, or labels.
  - Nominal: Categories do not have a natural ordering
    Eye colour; Nationality; Gender
  - Ordinal: Categories have a natural ordering
    Frequency of drinking Coke; Degree level

Note: We have no concept of distance with ordinal data.

Types of data



The type of data is important in determining which methods of analysis will be appropriate (and valid).

Types of Data

What data types do the following have?

- Number of students at AIMS
- Types of trees found in Cameroon
- Time taken to run 100m
- App ratings
- Comfortability level of a bed
- Age in years

Warning!

- The distinction between ordinal and numerical data can be difficult to determine in some cases.
- Consider App ratings. These usually take values 1,2,3,4,5, but could be argued that they are categorical values in disguise. *E.g.*
  - 1 – Very Dissatisfied
  - 2 – Dissatisfied
  - 3 – No opinion
  - 4 – Satisfied
  - 5 – Very Satisfied
- This is called a Likert scale.
- Commonly used, especially in surveys.

Warning 2!

- The distinction between discrete and continuous numerical data can be ambiguous too:
  - Suppose the area of a piece of land is recorded to the nearest m^2.
  - Or the weight of a person is recorded to the nearest tenth of kg.
- It could be argued that, although weight/land area are continuous, they are recorded in a discrete manner.
- However, it is always possible to compare two weights or two land areas, and say which is greater than the other. The chances of them being identical is infinitesimally small.
- For discrete data, instead, there are actual chances of two records being identical.

Descriptive Statistics

We describe data mainly via: **graphical** and **numerical** summaries.

Graphical Summaries

- Used for both qualitative and quantitative data.
- Provide a pictorial summary of the data
  - Barchart: Shows the distribution of qualitative data.
  - Histogram: Shows the distribution of quantitative data.
  - Scatter plot: Shows the relationship between two quantitative variables.

Numerical Summaries

- Only used for quantitative data.
- There are many numerical summaries, but the two main types are:
  - Measures of where the data is "centered"
  - Measures of "how spread" the data is

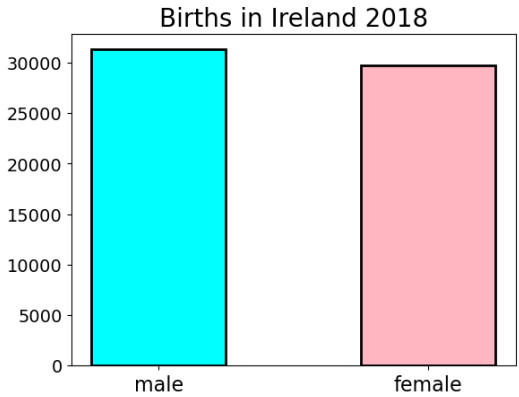
Graphical Summaries for Categorical Data

Barcharts (Example: Births)

- There were 61022 children born in Ireland in 2018.
- How did it break down into Male/Female births?

Males: 31,306

Females: 29,716



Barcharts

- Barcharts provide a graphical summary of **categorical data**.
- Very easy to produce:
  1. Count the number of observations in each category (**absolute frequency** F_i of each category).
  2. Draw a plot with a bar for each category: bars have equal width, and height proportional to absolute frequency.

Note: The height of each bar could also be the **relative frequency** f_i:

$$f_i = \frac{\text{Abs. Freq.}}{N} = \frac{F_i}{N}$$

where N is the total number of observations (in the population/sample).

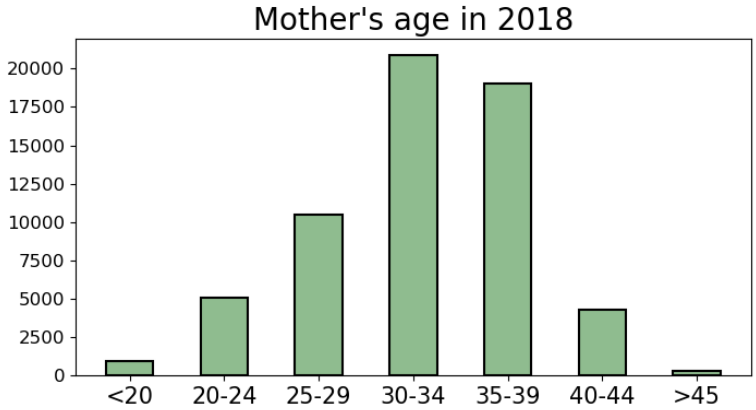
Barchats: Nominal and Ordinal Data

- Barchats are used for all categorical data (nominal and ordinal).
- The births example concerns `xxxxxxxnominal` data.
- But they can also be used for `ordinal` data.

Warning: If your data is ordinal, make sure that the bars are displayed in the right order.

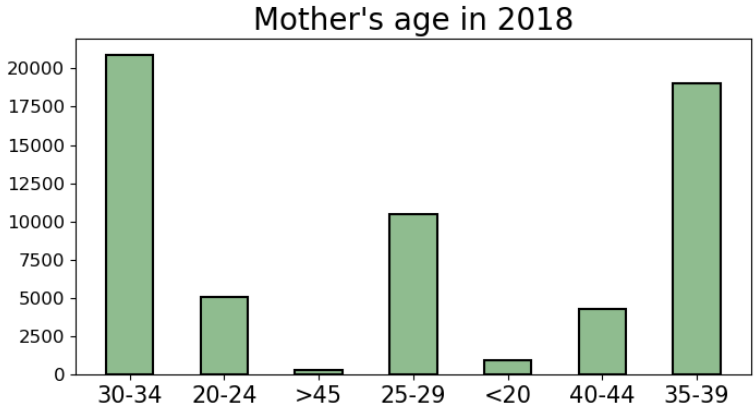
Example: Mother's Age

- Age of the mother when giving birth, coded into categorical bands.



Example: Mother's Age (not so good)

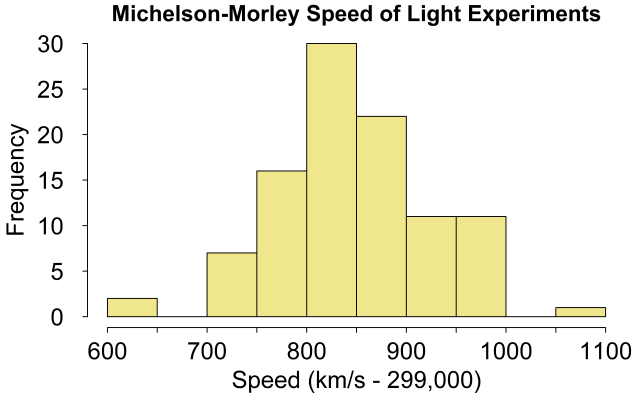
- The plot below is technically equivalent to the previous one, but it's not as clear and informative.



Graphical Summaries for Numerical Data

Histogram (Example: Speed of Light)

In 1887 Michelson and Morley conducted 100 experiments to measure the speed of light. We can represent the 100 measurements via a histogram:



Histogram

Histograms are used to represent numerical data.

1. Divide the range of the data into bins.
   Note: The bins are usually, but not necessarily, equal width.
2. Count how many observations fall into each bin.
3. Plot a rectangle for each bin, where the **area** of the rectangle is proportional to how many observations fall into that bin.

Let's make an example!

Producing a Histogram

A numerical sample contains:

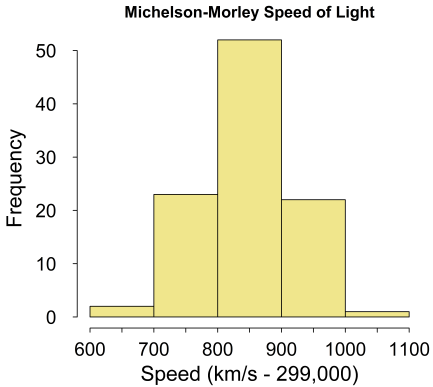
- 5 values between 0 and 2
- 7 values between 2 and 4
- 6 values between 4 and 7
- 4 values between 7 and 8

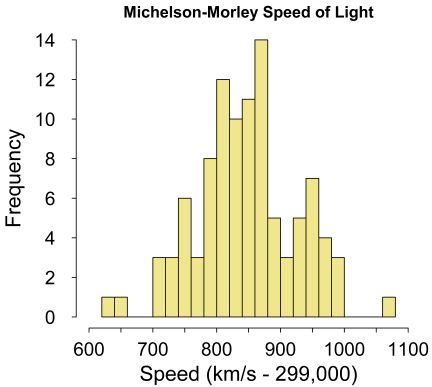
Draw a histoghram.

What happens if you put some of the bins together?

Example: Changing Bin Number

- The number of bins in a histogram can change its appearance.

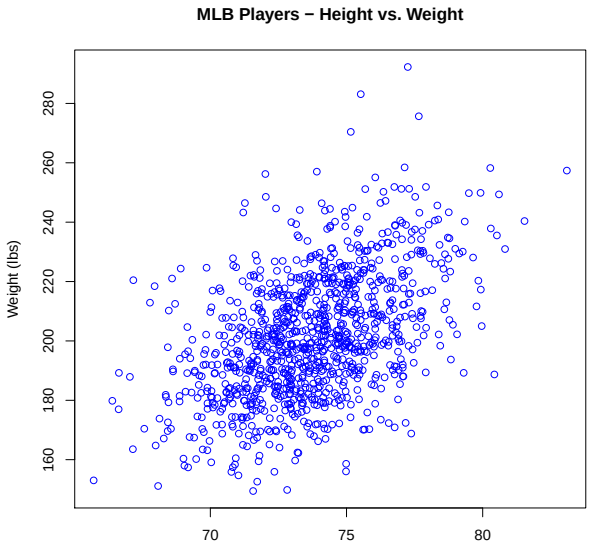




Relating two Variables: Scatterplots

- If we have measurements on two variables (*e.g.*, height and weight of players), plotting one variable against the other can be useful for finding patterns.
- **Scatterplots** are used for this: they provide an immediate visualisation of the relationship between two variables.
- They will be *very* useful in regression analysis (later in the course).

Scatterplot Example



Numerical Summaries for Quantitative Data

Numerical Summaries

- We have two main types of numerical summaries:
  1. Measures of location (mean, median, mode,...)
  2. Measures of spread (standard deviation/variance)
- Measures of location (or central tendency) give an idea of where the data are "centered".
- Measures of spread give an idea of the range of "most" of the data.

Measures of Location: Mean

- The **mean** or **average** is a measure of central tendency.
- Suppose we have **a sample of size** n from **a population of size** N:

$$x_i : \text{ value of the } i^{\text{th}} \text{ element}$$

- The **population mean** is

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

- The **sample mean** is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Mean from Aggregate Data

- When only a few different values are present, it is convenient to tabulate them with their frequencies, rather than reporting every single observation.
- For example, Age of AIMS students in 2025:

Age	Abs. Freq. (F_i)
22	5
23	11
24	23
25	17
26	27
27	18
28	7

Mean from Aggregate Data

In this case we can compute the mean as

$$\overline{x} = \frac{\sum_{i=1}^{K} F_i\, x_i}{\sum_{i=1}^{K} F_i} = \frac{\sum_{i=1}^{K} F_i\, x_i}{n} = \sum_{i=1}^{K} f_i\, x_i$$

where

- K in the total number of groups
- F_i is the absolute frequency of the value x_i.

Measures of Location: Median

The **median** M is the value that splits the data into two equal halves:

- 50% of the observations lie below it
- 50% of the observations lie above it.

To compute the median, first sort the data in increasing order. Then:

- If n is $\mathbf{odd}$: the median is the value at position $\frac{n+1}{2}$ in the sorted list.
- if n is $\mathbf{even}$: the median is the average of the two values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$ in the sorted list.

Median Examples

What's the median of the following sample: 6, 3, 12, 8, 9 ?

- Order the data: 3, 6, **8**, 9, 12
- Take the value in the middle:
  since $n = 5$ is odd, we take the element in position $\frac{(n+1)}{2} = 3$

$$\text{median} = 8\,.$$

What's the median of the following sample: 5, 3, 11, 8, 15, 12 ?

- Order the data: 3, 5, **8, 11**, 12, 15
- Take the value "in the middle":
  $n = 6$ even $\rightarrow$ take average of $\frac{n}{2} = 3\text{rd}$ and $\frac{n}{2}+1 = 4\text{th}$ elements

$$\text{median} = \frac{8+11}{2} = 9.5\,.$$

Measures of Location: Quartiles

The **quartiles** split the data into four equal quarters:

- The Lower Quartile Q_1 has 25% of the observations below it (and 75% above).
- The Middle Quartile Q_2 corresponds to the median.
- The Upper Quartile Q_3 has 75% of the observations below it (and 25% above).

To compute the lower and upper quartiles:

- Sort the data in increasing order.
- Find the values which lie 25% and 75% of the way through.

Quartiles Illustration

Let's visualise the ordered sample: each bar below represents one value.

Essentially: after finding the median M, we define:

- Q_1 as the median of the lower half
- Q_3 as the median of the upper half.

Example: Blood clotting time (Mean)

The following data report the blood clotting time (in seconds) for a sample of 14 subjects after administration of a particular reagent supposed to reduce the coagulation time.

27.87	29.34	28.35	29.24	29.51	29.48	28.84
29.28	29.34	29.10	28.51	28.52	28.36	29.46

- Mean:

$$\bar{x} = \frac{27.87 + 29.34 + \ldots + 29.46}{14} = 28.94\,.$$

Example: Blood clotting time (Median and Quartiles)

To compute Median, Lower and Upper quartile, let's re-order the data first:

27.87	28.35	28.36	**28.51**	28.52	28.84	**29.10**
29.24	29.28	29.34	**29.34**	29.46	29.48	29.51

- Median: As $n = 14$ even, we consider the average of the 7^{th} and 8^{th} values: (positions $\frac{n}{2}$ and $\frac{n}{2}+1$).

$$M = \frac{29.10 + 29.24}{2} = 29.17\,.$$

- Quartiles: For the quartiles, we find the medians of each of the two halves:

$$Q_1 = 28.51 \qquad Q_3 = 29.34\,.$$

Measures of Location: Mode

The **mode** is the most common value in the sample/population.

$$4, 2, 2, 5, 4, 3, 2, 5, 1 \quad \longrightarrow \quad \text{mode} = 2$$

- Note that:
  - The concept of mode applies to both categorical and numerical data
  - However, for numerical data (especially continuous), the mode may not be well defined.
  - It is possible to have more than one mode.

- The sample mean, sample median and sample mode are examples of **statistics**: numbers calculated from the data which, in some way, summarise the data.

Remarks

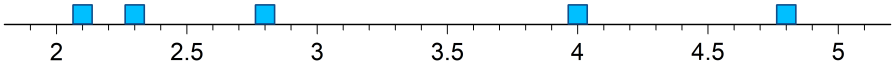
- The mean can only be computed for **numerical** data (discrete and continuous)
- The median can be computed for **numerical and ordinal** data
- The mode can be computed for **all types** of data (at least in principle, see previous slide)

Mean: Visualisation

Consider the five values:

$$2.8 \quad 2.1 \quad 4.8 \quad 2.3 \quad 4\,.$$

Imagine each data point as a small weight on a horizontal bar.



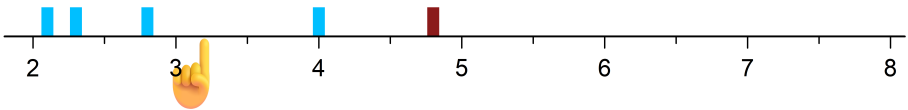
Where would you hold the bar from below so that it stays balanced?

That spot is the mean of the data!

Mean vs Median: Visualisation

Same values as before (2.8, 2.1, 4.8, 2.3, 4).

Suppose the value **4.8** ...



...turned out to be a **7.3** instead. How do the sample mean/median change?

- Mean $\rightarrow$ very sensitive to outliers (can shift a lot).
- Median $\rightarrow$ robust to outliers (barely affected).

Sensitivity to Outliers: Example

- Compute the mean and median for the sample:

$$6.5,\ 10,\ 3,\ 7.5,\ 8$$

Now replace the 8 with 78. How have mean and median changed?

- Compute the mean and median for the sample:

$$10,\ 5,\ 7,\ 6,\ 3$$

Replace the number 5 with 50, and check what has changed.

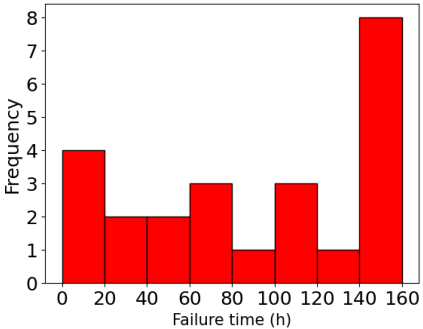
Example: Failure Times

Twenty-four pumps were tested until they failed. The time to failure (in hours) was recorded.

6.0	8.6	17.8	18.0	27.5	33.5	50.5	51.5
69.0	74.0	74.0	89.0	109.0	118.0	119.0	138.0
141.0	144.0	146.0	150.0	151.0	153.0	153.1	153.2

How can we summarise the data?

- Graphical Summary: Histogram
- Numerical Summary:
  - Mean: 91.45 h
  - Median: 99 h



Measures of Spread (Variability)

Measuring Variability for Quantitative Variables

Ways to measure the spread of a numerical sample $x_1, x_2, \ldots, x_n$.

- **Range**: Measures the full spread of the data.
- **Interquartile Range (IQR)**: Measures the spread of the central half of the data.
- **Variance** and **standard deviation**: Measure how far values are from the mean.

Measures of Spread: Range

- The **range** records the difference between the maximum and the minimum value in the data.
- It records the full range of the data. However, it is *highly sensitive* to outliers (extreme values in the data).
- The range has the same "units" as the data itself (clearly).

Measures of Spread: Interquartile Range

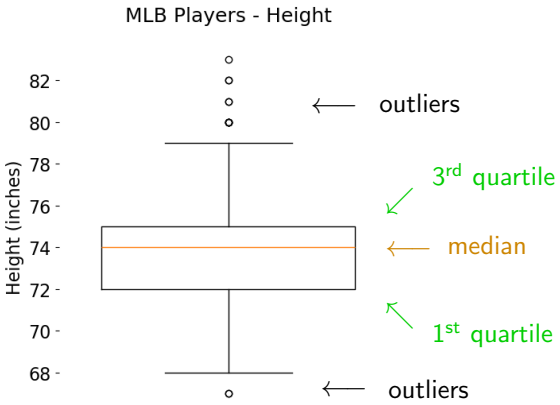
- The **IQR** is the difference between the upper and lower quartiles:

$$\text{IQR} = Q_3 - Q_1 \,.$$

- It records the range of the central 50% of the data.
- It is very robust to outliers.
- Also the IQR has the same "units" as the data (as do Q_1 and Q_3).

Box Plots

- Box plots provide a useful graphical summary of the data.
- They report where the first, second and third quartiles of the data lie, as well as showing the more "extreme" observations.



Measures of Spread: Variance and Standard Deviation

Remember, we start from a numerical sample:

$$x_1, x_2, \ldots, x_n\,.$$

The variance and standard deviation measure how far the x_i are from the mean.

The formula for the **variance** is:

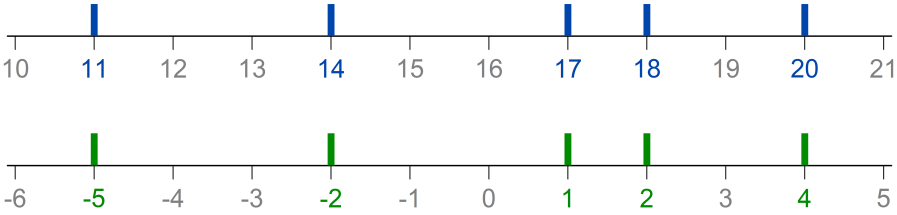
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

What does that mean? Let's go through an example.

Variance: Graphical Meaning/Intuition

Consider the following sample of size $n = 5$:

x_i	11	14	17	18	20	($\bar{x} = 16$)
$x_i - \bar{x}$	-5	-2	1	2	4	(mean $= 0$)
$(x_i - \bar{x})^2$	25	4	1	4	16	(mean $= 10$)



Variance and Standard Deviation

The **variance** of a sample is computed as the "average" squared difference between each x_i and the sample mean $\bar{x}$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

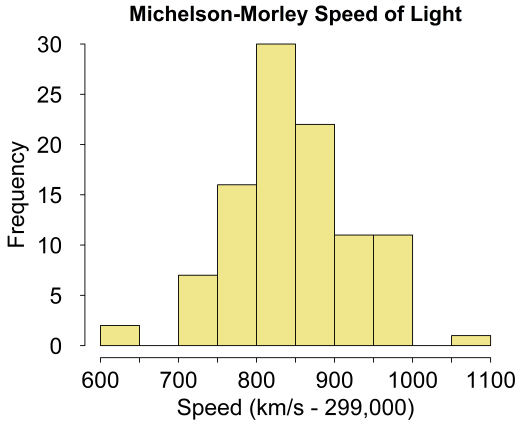
The **standard deviation** is the square root of the variance $s = \sqrt{s^2}$.

Variance and Standard Deviation

- The **variance** has the **squared units** of the data
- The **standard deviation** (SD) has the **same units** of the data.
- Both of them are *not* robust to outliers.

Measures of Spread: Standard Deviation

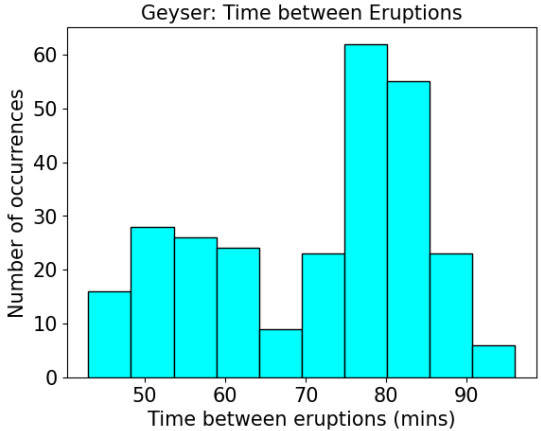
Larger values of the SD indicate greater spread of data.



$\longrightarrow$ SD = 79

Measures of Spread: Standard Deviation

Larger values of the SD indicate greater spread of data.



$\longrightarrow$ SD $= 13.6$

Percentiles

- We have seen the definition of the 25^{th} and 75^{th} percentiles (first and third quartile, respectively).
- However, **percentiles** can be of any order between 0 and 100.
- They identify the value below which the specified percentage of a sample (or population) falls. For example:
  - The 1^{st} percentile of a sample is the value such that 1% of the sample's values fall below it, and 99% fall above it
    (so, it's one of the smallest values in the sample)
  - The 90^{th} percentile will have 90% of values below it, and 10% of values above it
    (a fairly large value within the sample).

Using Summaries in Conjunction

In most situations, we find that "most" of the data lie in the range

$$\text{Median} \pm 1.5 \times \text{IQR}$$

and/or

$$\text{Mean} \pm 2 \times \text{SD}$$

- Later in the course, we will provide mathematical reasons for why, in many cases, 95% of the data lie in the **second range**.

Recap of this Unit

What we have seen in this unit:

- **How to classify data**
  - Quantitative (numerical) & Qualitative (categorical)
  - Discrete/Continuous & Nominal/Ordinal
- **Graphical ways to present data**
  - Barcharts (categorical)
  - Histograms, Boxplots (quantitative)
  - Scatterplots (two quantitative)
- **Numerical summaries of data**
  - Measures of location (mean, median, mode, quartiles/percentiles)
  - Measures of spread (range, IQR, variance, standard deviation)

Next, we move to an introduction to **Probability**.